NGS Modeling

NGS Modeling

New models need to be developed and implemented to accommodate NGS data, with the ultimate goal of developing a probabilistic approach for NGS mixture interpretation.

CE-based models can be used as a basis for NGS modeling. Both methods make use of the PCR process, so it is expected that artifacts such as stutter are similar.

However, peak heights need to be substituted with read counts and the remaining biological processes differ. This will materially affect the modeling parameters.

NGS Stutter Modeling

NGS data generally show higher stutter percentages than CE data. Illumina's ForenSeq uses the following thresholds (compared with Thermo Fisher's NGM Select Kit for CE data):

	Stutter Filter (%)	
Locus	**CE**	**NGS**
TH01	5	10
D2S441	9	7.5
vWA	11	22
FGA	11.5	25
D12S391	15	33
D22S1045	17	20

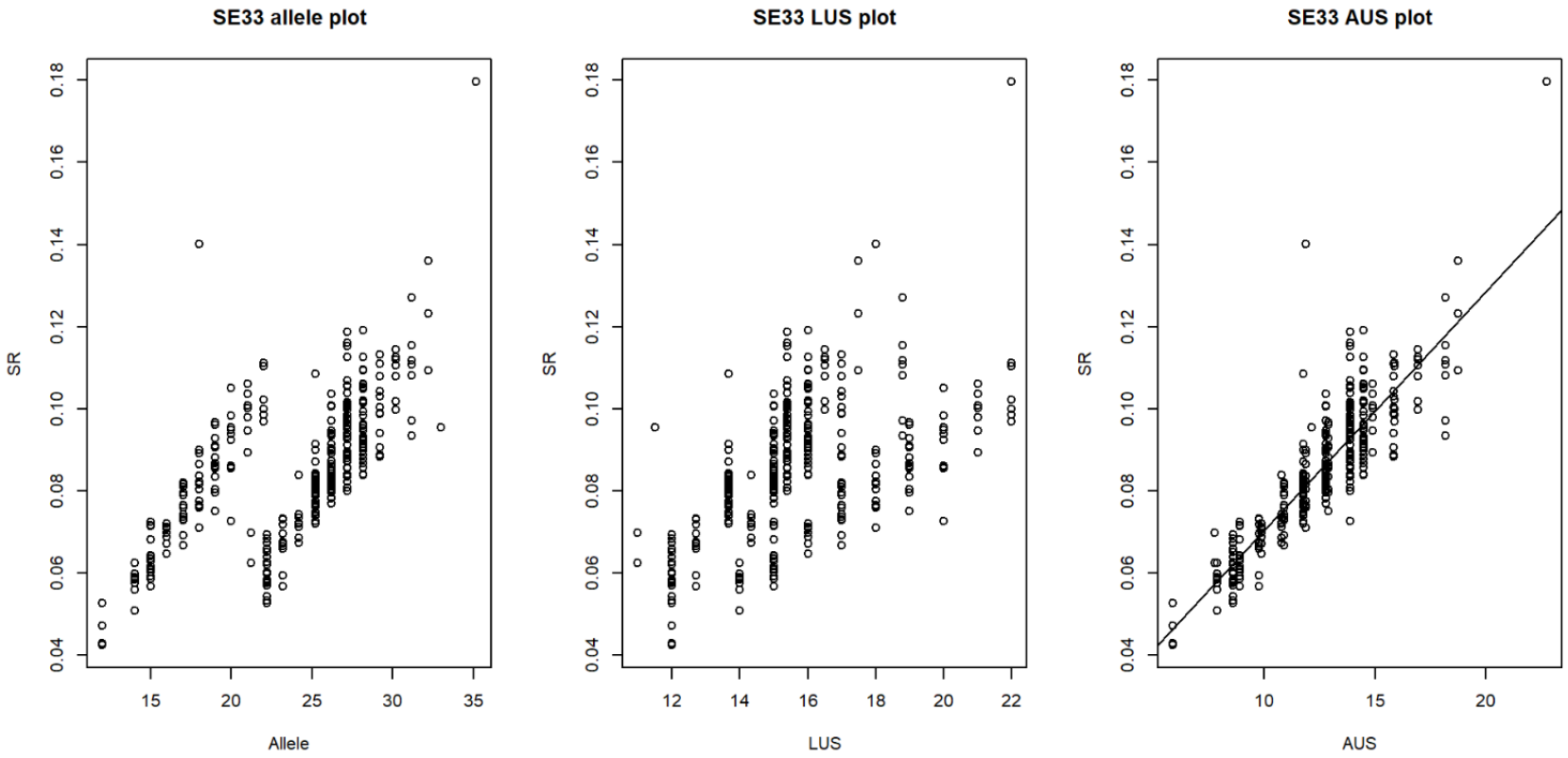
Multi-sequence Stutter Model

A multi-sequence model takes into account all uninterrupted stretches (AUS) as potentially contributing to stuttering.

Allele	**Repeat motif**
21.2	$[AAAG]_2AG[AAAG]_3AG[AAAG]_9AA\ AAAG[AAAG]_{11}G\ AAGG[AAAG]_2AG$
21.2	$[AAAG]_2AG[AAAG]_3AG[AAAG]_{11}AA\ AAAG[AAAG]_9G\ AAGG[AAAG]_2AG$
22	$[AAAG]_2AG[AAAG]_3AG[AAAG]_{22}G[AAAG]_3AG$
22.2	$[AAAG]_2AG[AAAG]_3AG[AAAG]_7AA\ AAAG[AAAG]_{14}GAAGG[AAAG]_2AG$
22.2	$[AAAG]_2AG[AAAG]_3AG[AAAG]_8[AG]_5[AAAG]_{12}GAAGG[AAAG]_2AG$
22.2	$[AAAG]_2AG[AAAG]_3AG[AAAG]_9AA\ AAAG[AAAG]_{12}GAAGG[AAAG]_2AG$

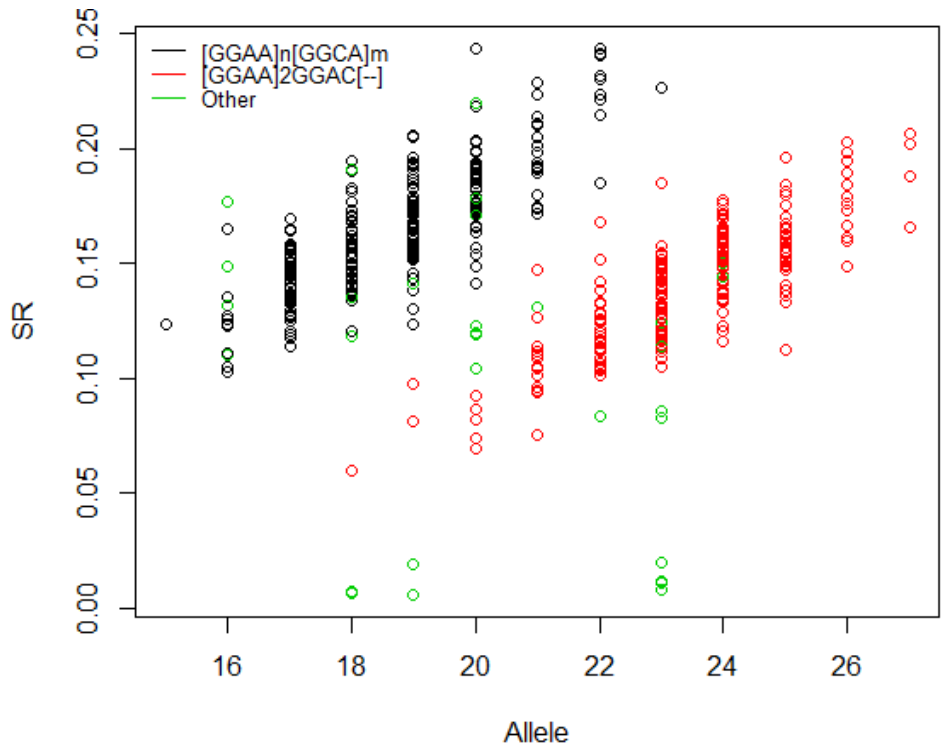
Examples of locus SE33 sequences.

Multi-sequence Stutter Model for SE33



Stutter Modeling and Sequence Variation

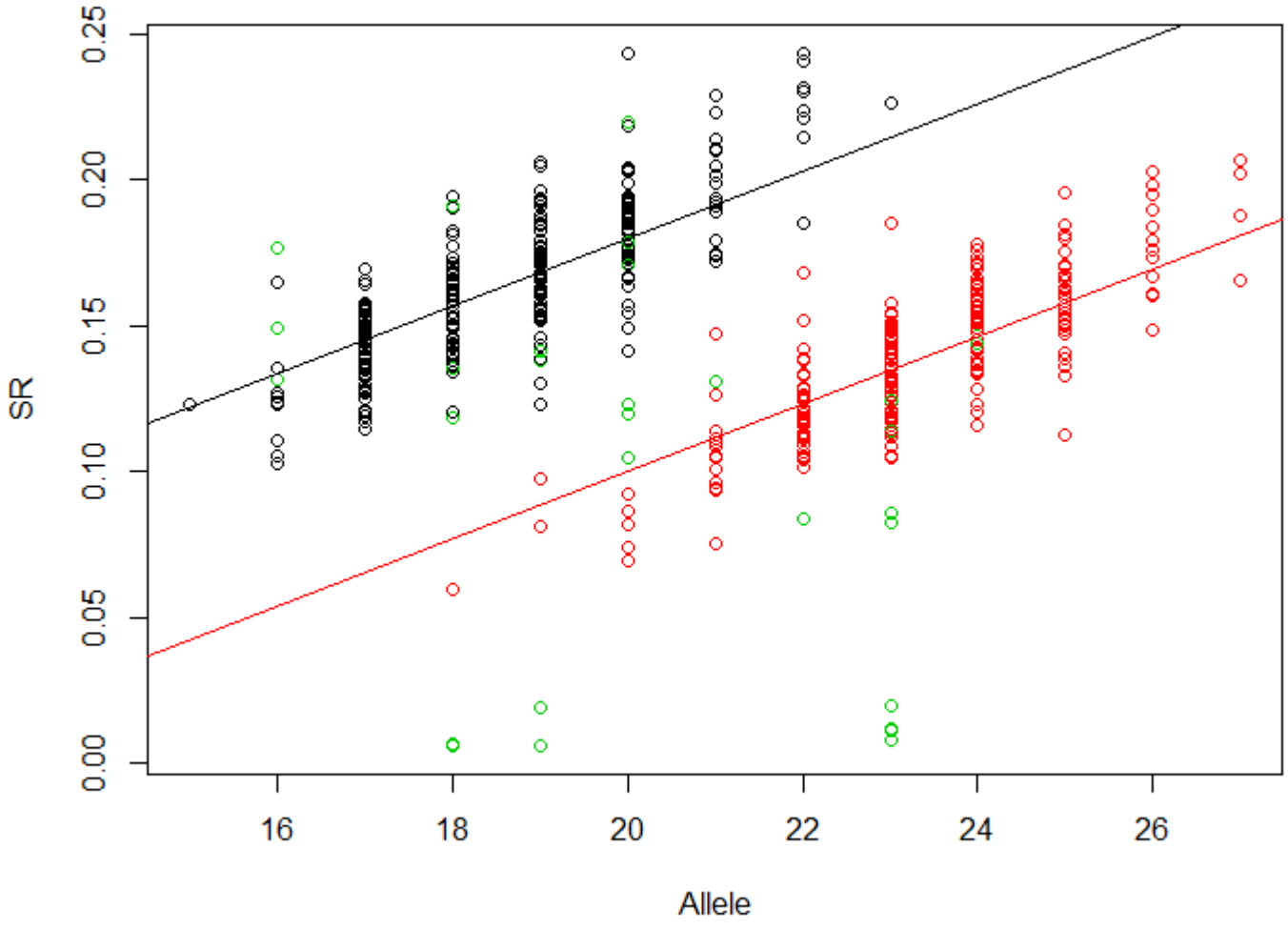
What about variation that is suggested to be attributable to sequence motif?



Stutter ratios for locus D2S1338.

Models fitted based on AUS still left some variability unexplained for some loci.

NGS Stutter Modeling - Sequence Variation



Stutter ratio model for locus D2S1338.

NGS Stutter Modeling

With the sequence variations now in hand, it is possible to decompose certain stutter affected heterozygotes, composite stutter and regular stutter products.

For locus TH01, for example, there are two possible (back) stutter products:

Product	**LB Allele**	**SB Allele**
A	8.3	$[\mathrm{AATG}]_6\mathrm{ATG}[\mathrm{AATG}]_2$
B	8.3	$[\mathrm{AATG}]_5\mathrm{ATG}[\mathrm{AATG}]_3$

NGS Stutter Modeling

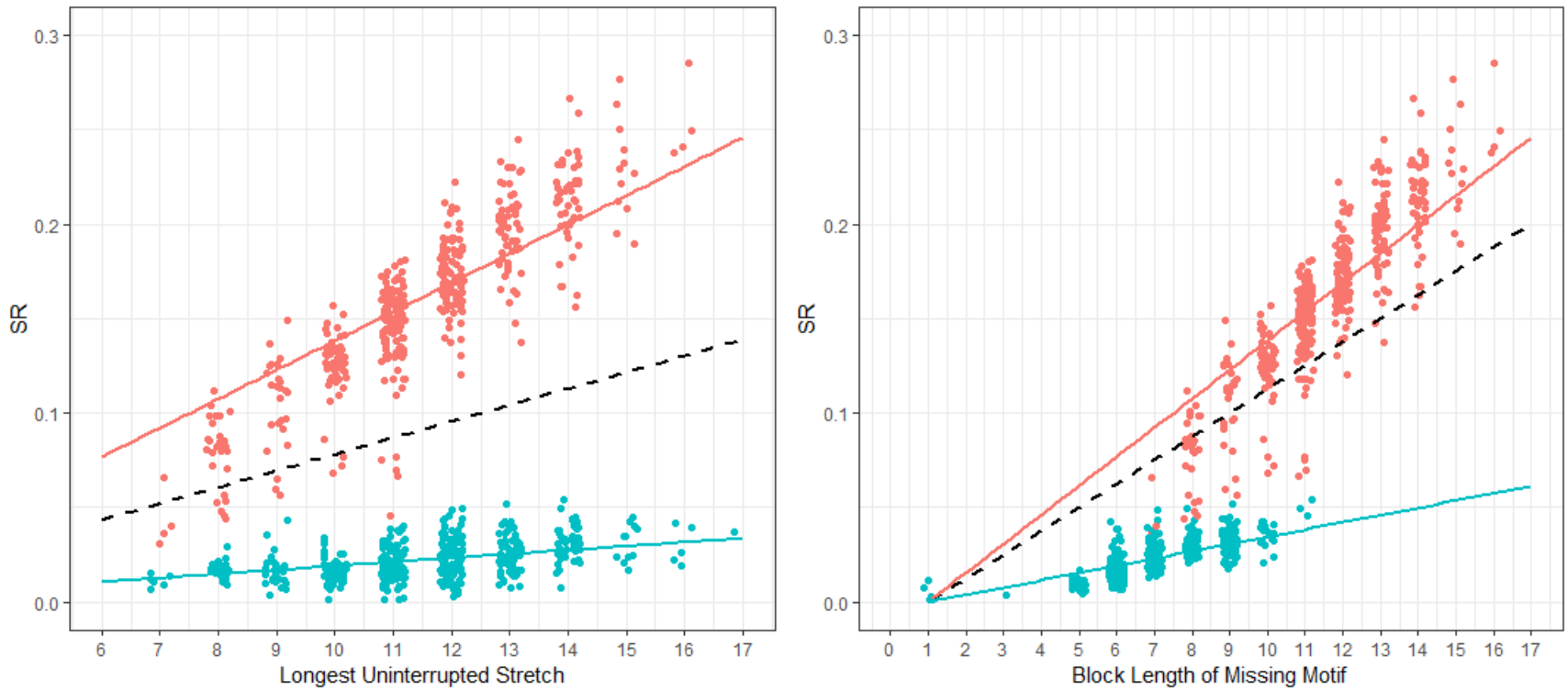
Recall the definition of the stutter ratio:

$$SR = \frac{O_{a-1}}{O_a} = \frac{O_A + O_B}{O_a} = \frac{O_A}{O_a} + \frac{O_B}{O_a}$$

Instead of modeling stutter per parental allele, you can also model the ratios per different stutter sequence. This was not possible for CE data.

Category	**Allele**	**Sequence**	**Count**	**SR**
Allele	9.3	$[AATG]_6ATG[AATG]_3$	100	0.25
Stutter	8.3	$[AATG]_6ATG[AATG]_2$	5	0.05
Stutter	8.3	$[AATG]_5ATG[AATG]_3$	20	0.20

NGS Stutter Modeling



Stutter ratio model for locus D12S391.

The larger stutter ratios result from stutter from the LUS of the parental allele.

NGS Stutter Modeling - Discussion

- How to determine variation?
- What about micro-variants?
- What about the possible influence from flanking variation?
- What about dependencies between stretches?

NGS Population Structure

Likelihood ratios use match probabilities, which rely on appropriate estimation of the population structure parameter θ. Values of around 3% are common in forensic DNA evidence evaluations.

When implementing NGS-based methods, the effect of sequence data on θ estimates needs to be analyzed.

Allele and/or genotype matching between individuals within and between populations can help us assess relative relatedness[1].

[1] Population-specific F_{ST} values for forensic STR markers: A worldwide survey (Buckleton et al., 2016).

NGS Population Structure

Our data consist of DNA samples from 350 individuals over 5 different continental groups sequenced and annotated with Illumina instrumentation.

- Using length-based allele callings, within-population matching was 0.2165 and between-population matching was 0.1968.

- Using sequence-based allele callings, within-population matching was 0.1878 and between-population matching was 0.1664.

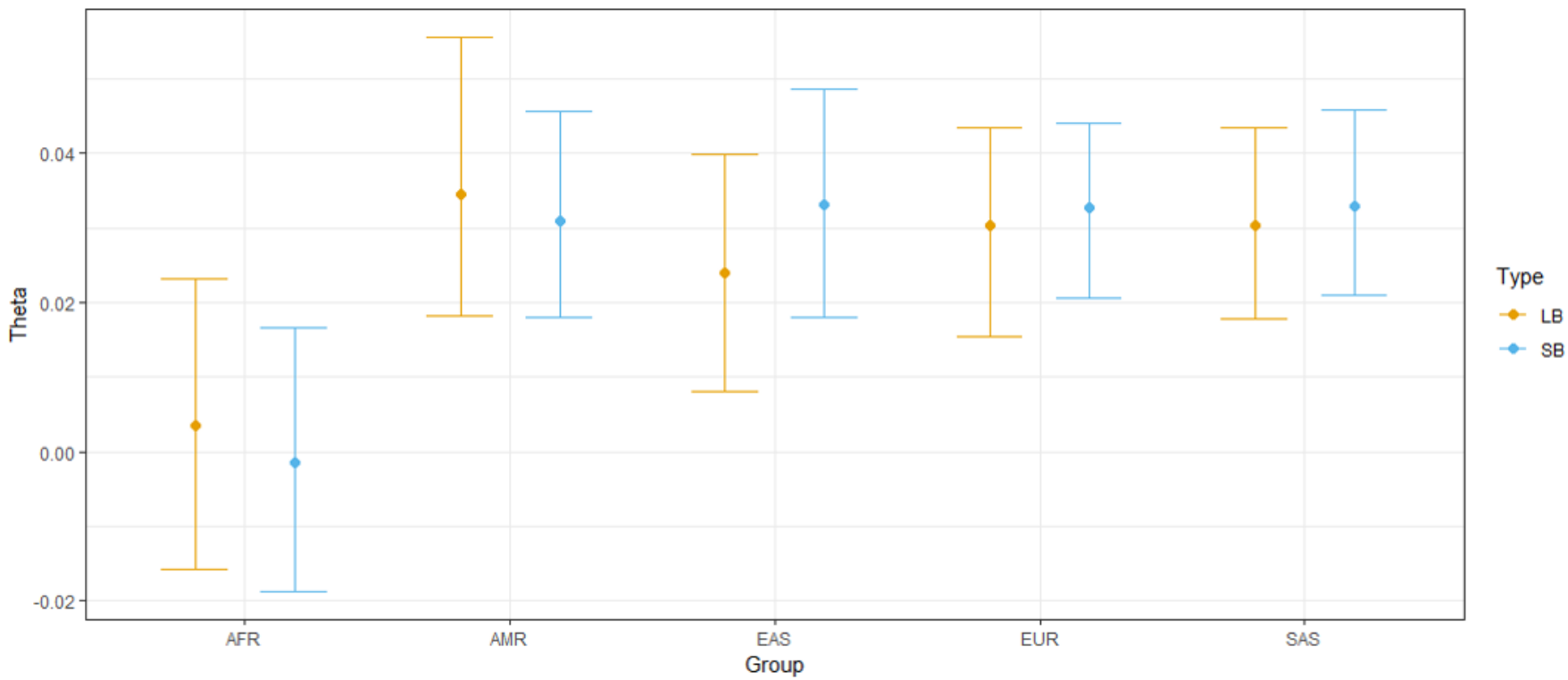
NGS Population Structure

Locus-specific θ estimates may decrease, increase, or stay the same.

	# Alleles			θ Estimate	
Locus	**LB**	**SB**	**Diff**	**LB**	**SB**
D21S11	17	65	48	0.0259	0.0383
D1S1656	18	34	16	0.0174	0.0146
TPOX	8	8	0	0.0402	0.0402

NGS Population Structure

Results show very similar effects of sequencing data on theta estimates as what we have seen for CE-based results.



Confidence intervals per group.